IJDC | *Conference Pre-print*

Extending Support for Publishing Sensitive Research Data at the University of Bristol

Zosia Beckles
University of Bristol

Abstract

The University of Bristol Research Data Service was set up in 2014 to provide support and training for academic staff and postgraduate researchers in all aspects of research data management. As part of this, the data.bris Research Data Repository was developed to provide a publication platform for research data generated at the University of Bristol. Initially launched in 2015 to provide open access to data, since 2017 it has also been possible to publish access-controlled datasets containing sensitive data via this platform.

The vast majority (90%) of datasets published are openly accessible, but there has been steady demand for access-controlled release of datasets containing information that is ethically or commercially sensitive. These cases require careful management of additional risk: for example, where datasets contain information on human participants, balancing the risk of re-identification with the need to provide robust data that maximises research value through re-use. Many groups within the University of Bristol (for example, the Avon Longitudinal Study of Parents and Children) have extensive experience and expertise in this area, but it became apparent that there was a need to provide additional support for researchers who were not able to draw on the experience of these established groups. This practice paper describes the process of setting up a dedicated service to provide training and basic disclosure risk assessments in order to address these skills gaps, and outlines lessons learnt and future directions for the service.

Submitted 15 December 2019 ~ *Accepted* 19 February 2020

Correspondence should be addressed to Zosia Beckles, University of Bristol, Beacon House, Queens Road, Bristol, BS8 1QU, UK. Email: z.beckles@bristol.ac.uk

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The University of Bristol Research Data Service was set up in 2014 to provide support and training for academic staff and postgraduate researchers in all aspects of research data management. As part of this, the data.bris Research Data Repository was developed to provide a publication platform for research data generated at the University of Bristol. Initially launched in 2015 to provide open access to data, since 2017 it has also been possible to publish access-controlled datasets containing sensitive data via this platform.

The vast majority (90%) of datasets published are openly accessible, but there has been steady demand for access-controlled release of datasets containing information that is ethically or commercially sensitive. These cases require careful management of additional risk: for example, where datasets contain information on human participants, balancing the risk of re-identification with the need to provide robust data that maximises research value through re-use. Many groups within the University of Bristol (for example, the Avon Longitudinal Study of Parents and Children) have extensive experience and expertise in this area, but it became apparent that there was a need to provide additional support for researchers who were not able to draw on the experience of these established groups. This practice paper describes the process of setting up a dedicated service to provide training and basic disclosure risk assessments in order to address these skills gaps, and outlines lessons learnt and future directions for the service.

Data publication at the University of Bristol

Original process

Research Data Service staff carry out pre-publication checks on all datasets, including those containing potentially sensitive data. However, Research Data Repository users are expected to take responsibility for preparing their data, including ensuring that appropriate consent is in place for data sharing and that sensitive data is appropriately anonymised.

For the most part, processes for publishing open and access-controlled data are similar, and the following pre-publication checks carried out by the Research Data Service for all datasets:

1. Dataset complies with data preparation rules[1]
2. Dataset contents match file inventory
3. File functionality (for large datasets, checks may be carried out on a sample only)
4. Third party data licensed appropriately
5. Correct/appropriate metadata in deposit record

For controlled-access datasets, additional checks are carried out in these areas:

6. Consent form and patient information sheet included in dataset, if applicable
7. Dataset access level is appropriate to contents and, if applicable, matches terms of consent

[1] http://www.bristol.ac.uk/staff/researchers/data/publishing-research-data/data-preparation-rules/

Skills gaps

As noted previously, the onus is placed on data depositors to ensure that dataset release risks (for example, participant re-identification) have been adequately controlled. For commercially sensitive data, researchers were generally confident in making decisions on the appropriate level of access, and could access expert legal advice from the University of Bristol Secretary's Office or Research and Enterprise Development Contracts team. However, ethically sensitive data containing information on human participants made up the majority of controlled-access data publications, and was often deposited by researchers with no prior experience in publishing this type of sensitive data. These requests were often further complicated by ambiguous or missing consent for data sharing, requiring approval to be sought after the fact from the original Research Ethics Committee, or the dedicated University of Bristol Data Access Committee.

The Research Data Service already had some guidance on handling sensitive data[2], but an increasing number of requests were received for assistance beyond the scope of this, in particular for advice on appropriate anonymisation of spreadsheets and interview transcripts. Initially, these were referred to centres of expertise within the University, such as the Jean Golding Institute for data intensive research[3], and to general guidance available elsewhere such as the Anonymisation Decision-Making Framework[4] and the UK Data Service[5]. However, as demand increased it became clear that this ad hoc approach was not sustainable: preparing a single sensitive dataset for publication required extensive discussion with multiple experts around the university, and multiple iterations of the amended dataset before a balance was struck that met the requirements of the depositor and likely data re-users whilst protecting the privacy of the participants. It was therefore clear that additional ongoing in-house support was required to effectively deal with these complex cases, and to fill these gaps in skills and experience across the university.

New sensitive data support service

To meet this increased demand, a new 1.0 FTE Research Information Analyst (RIA) post was created in the Research Data Service to provide specialist support for publishing ethically sensitive datasets. The RIA post is a dual role, covering both sensitive data and research analytics; the sensitive data role will be the focus of this paper. Sensitive data support comprises two main areas:

1. disclosure risk assessments for datasets containing sensitive information
2. training and guidance for academics working with sensitive data

A disclosure risk assessment workflow was developed to extend the controlled-access pre-publication checks provided by the Research Data Service. When a request is received to publish a dataset containing information relating to human participants, this is passed to the RIA for additional checks, as shown in Figure 1.

[2] http://www.bristol.ac.uk/staff/researchers/data/dealing-with-sensitive-data/
[3] http://www.bristol.ac.uk/golding/
[4] https://ukanon.net/ukan-resources/ukan-decision-making-framework/
[5] https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation

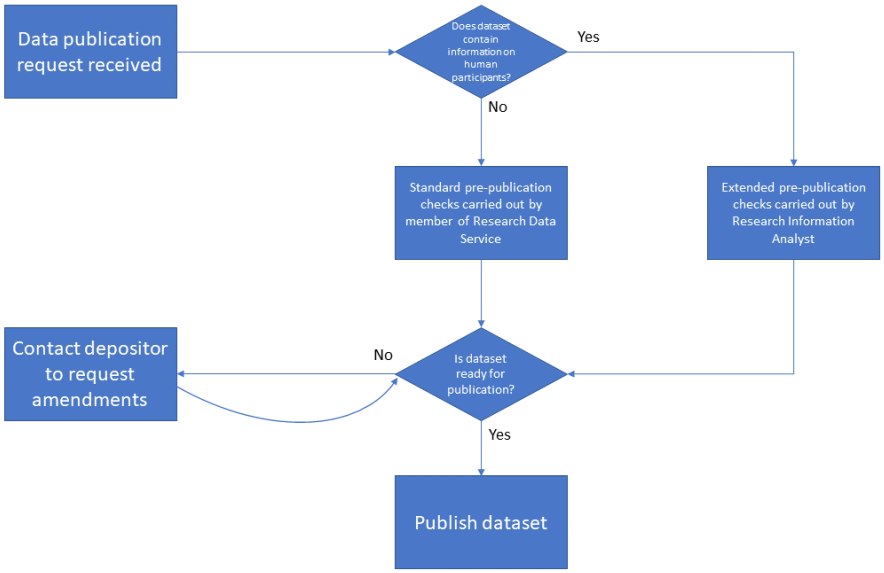

Figure 1: modified dataset publication workflow

The extended process includes checks for consent forms and patient information sheets, and for appropriateness of access level as outlined in the original data publication process above, but also uses tools such as the text anonymisation helper tool[6], QAMyData (both developed by the UK Data Service), and the sdcMicro[7] R package to investigate the data itself for disclosure risks as follows:

1. Assessment of the dataset key variables (for example age, gender, geographical location) and data environment for additional information contained in published protocols or accompanying papers

2. Checks for direct identifiers (manually or via QAMyData)

3. For quantitative datasets, assessment of disclosure risk in sdcMicro

If the risk of disclosure is considered unacceptable, the depositor is contacted with suggested amendments. No changes are made to the dataset without the depositor's approval. Suggested amendments may include formal anonymisation (removal of direct identifiers) or basic statistical anonymisation (for example, aggregating or grouping variables). If more complex statistical techniques are required (for example, adding noise to data), the depositor is referred to other centres of expertise within the University.

At the same time, new online training resources and workshops for academics were developed to address the skills gaps identified. A new three-hour workshop was developed, addressing the following topics:

- analysis of consent form wording and the impact on data sharing
- sources of disclosure risk

6 https://ukdataservice.ac.uk/media/622369/md5_94fc0c2a25f3a75396059826a23b8224_textanonymisationhelpertool_01_00.zip

7 https://cran.r-project.org/package=sdcMicro

- methods for managing disclosure risk, including formal and statistical anonymisation, and functional anonymisation via controlled-access data release
- legislative and regulatory requirements, including the General Data Protection Regulation and the Data Protection Act 2018

This workshop proved to be one of the most popular offered by the Research Data Service: fully booked on each occasion it has been delivered, and consistently receiving positive feedback from attendees. Due to demand will be repeated multiple times throughout the forthcoming academic year.

Online training resources (still in development) will cover similar areas, allowing greater uptake by those unable to attend live events. A series of video interviews were carried out with key stakeholders at the University of Bristol, including academics with experience of sharing sensitive data, members of the Research and Enterprise Development Research Governance team, and staff from the Secretary's Office with information governance responsibilities.

Outcomes and next steps

Since the launch of the sensitive data support service in April 2019, six access-controlled datasets have been published by the Research Data Service. All of these would have required the involvement of a number of different experts across the University, and prior to the launch of the service would have taken anywhere from a few weeks to a few months to resolve, dependent on the availability of these external staff. Instead, the ability to handle the majority of these cases in-house means that publication of a typical access-controlled dataset can be completed within two weeks; longer than for open datasets, but a significant improvement on the previous timescale.

The immediate future goal for sensitive data support is finalisation of the online training resources, including developing a new interactive 'bootcamp' tutorial. These resources will not replace the live workshops, but will provide an entry point for those unable to attend in-person training. Finally, outreach and advocacy activities have to date been reactive, based on incoming requests, so in the future a more proactive approach must be used to ensure that service uptake is widespread.

References

Hiom, D., Gray, S., Steer, D., Merrett, K., Snow, K., Beckles, Z. (2017). Introducing Safe Access to Sensitive Data at the University of Bristol. *International Journal of Digital Curation, 12(2)*. doi:10.2218/ijdc.v12i2.506